Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society.
Albuquerque, NM, Human Factors Society, 1997

CANONICAL CORRELATION ANALYSIS OF DATA ON HUMAN-AUTOMATION INTERACTION

Michael G. Shafto
NASA Ames Research Center
Moffett Field, CA

Asaf Degani
San Jose State University
San Jose, CA

Alex Kirlik
Georgia Institute of Technology
Atlanta, GA

ABSTRACT

Canonical correlation analysis is a type of multivariate linear statistical analysis, first described by Hotelling (1935), which is used in a wide range of disciplines to analyze the relationships between multiple independent and multiple dependent variables. We argue that canonical correlation analysis is the method of choice for use with many kinds of datasets encountered in human factors research, including field-study data, part-task and full-mission simulation data, and flight-recorder data.

Although canonical correlation analysis is documented in standard textbooks and is available in many statistical computing packages, there are some technical and interpretive problems which prevent its routine use by human factors practitioners. These include problems of computation, interpretation, statistical significance, and treatment of discrete variables. In this paper we discuss these problems and suggest solutions to them. We illustrate the problems and their solutions based on our experience in using canonical correlation in the analysis of a field study of crew-automation interaction in commercial aviation.

INTRODUCTION

Canonical correlation analysis (CCA) is a type of multivariate linear statistical analysis, first described by Hotelling (1935). It is currently being used in fields like chemistry, biology, meteorology, demography, artificial intelligence, cognitive science, political science, sociology, psychometrics, educational research, economics, and management science to analyze multidimensional relations between multiple independent and multiple dependent variables.

Although CCA is documented in standard textbooks, and is available in statistical computing packages, there are certain technical and interpretive problems which prevent its routine use by human factors practitioners. These include problems of computation (matrix singularity, computer time requirements), interpretation (visualization, examination of individual cases), and statistical significance (significance levels and confidence intervals for multidimensional non-normal data, including discrete variables).

In this paper we discuss these problems and some solutions to them. We illustrate the problems and their solutions based on our experience in using canonical correlation in the analysis of a field study of crew-automation interaction in commercial aviation. We begin with a brief description of CCA, followed by a description of our field study and the data we were analyzing. We describe five specific problems we encountered during the analysis, along with our proposed solutions to each problem. We conclude with an assessment of the utility of CCA in the context of a spectrum of analytical methods for complex, real-world data.

CANONICAL CORRELATION ANALYSIS (CCA)

CCA is an extension of multiple regression to the case of a multidimensional response measure. As explained by McKeon (1965), most of the familiar methods of linear statistical analysis, including multiple analysis of variance (MANOVA), discriminant analysis, and principal components analysis, are special cases of CCA.

CCA is most closely related to multiple regression and to principal components analysis. CCA differs from principal components analysis (PCA) in that PCA makes no distinction between independent and dependent variables. CCA differs from multiple regression in that CCA allows multiple dependent variables; multiple regression allows multiple independent variables but only a single dependent variable. Therefore, CCA is a potentially valuable tool in human factors research involving (1) a clear distinction between independent and dependent variables, (2) multiple dependent variables, and (3) the potential for multidimensional relations between these two sets of variables. For example, these conditions often apply to field studies of decision making and performance, field trials of product or system usability, part-task or full-mission simulation studies of expert performance, and on-line performance data such as flight recorder data.

Although the numerical methods required for CCA are much more complex than those required for computing a bivariate correlation coefficient, CCA can be conceptually understood in terms familiar from bivariate analysis (Cliff, 1987). CCA computes two derived (or “canonical”) variables, X and Y, such that the correlation between X and Y is as large as possible. X is a weighted average of the independent variables, and Y is a weighted average of the dependent variables. CCA computes several such relations between independent and dependent variables. Each relation indicates a distinct pattern that exists in the data. CCA reduces each of these patterns to a bivariate correlation between just two derived variables, the canonical X and Y variables. This means, for example, that each such relation can be visually inspected using a familiar bivariate scatter diagram.

CCA extends the concept of bivariate correlation in one important way: There may be several patterns of correlation between the independent and dependent variables, each pattern corresponding to a different pair of weighted averages. Hence, the results of CCA are usually reported in terms of two lists of weights (one for X and one for Y) and a canonical correlation for each pattern. The largest canonical correlation corresponds to the strongest relation between independent and dependent variables; subsequent canonical correlations correspond to relations of decreasing strength. The significance of this feature for human factors research is that we often find different response patterns under different environmental conditions, for example, different pattern of autoflight mode selection under different phases of flight. CCA allows these patterns to be characterized objectively, and allows their relative strengths to be measured. Further, by using the bivariate scatter diagram mentioned above, CCA allows the analyst to focus his or her attention on outliers which fail to conform to the dominant patterns; for example, crews flying manually rather than using fully automatic modes during the en route flight phase.

FIELD STUDY AND DATASET.

The study involved cockpit observations of crew interaction with the automatic flight control system of the Boeing 757/767 aircraft during revenue flights by a major U.S. air carrier. Each data record characterized a change in mode selection, along with a number of variables describing the conditions under which the change occurred. The initial dataset used in the CCA analysis consisted of over 1500 records, each characterized by 75 variables. Approximately half the variables had to do with the context or situation, and the other half had to do with the crew's response, that is, their choice of an autoflight mode configuration. A full description of the field study may be found in Degani (1996).

Initial Questions

In general, we were interested in characterizing relations between situations and response patterns, that is, between the state of the operating environment and human action (mode selection). The value of using CCA in this case derived from its unique suitability for finding patterns in large datasets. We had multiple independent variables which characterized operational situations (ATC clearance given, captain vs. first officer flying, distance from airport, altitude, ATC facility, clearance given, departure and destination airports), as well as multiple dependent variables consisting mainly of categorical variables used to describe the crew's selection of autoflight modes. In addition to characterizing patterns of situation-response relations, we wanted to be able to recognize unusual cases (outliers), in order to focus our analysis on those individual cases which might illuminate unusual crew behavior or crew error. Finally, by using CCA for this initial data-reduction analysis, we used both the typical behavior patterns and atypical cases (outliers) as points of departure for developing dynamic models of crew-automation interaction (Degani & Kirlik, 1995).

PROBLEMS AND PROPOSED SOLUTIONS

Redundant variables

The first problem we encountered was that the correlation matrix for the original variables was singular. This is a common problem when the number of variables is large—it simply means that some of the variables are redundant. It is difficult, however, to determine by inspection alone which variables are redundant. We used several methods to attack this problem, but the best overall solution was to use cluster analysis prior to the CCA.

Cluster analysis (Jardine & Sibson, 1971) is a way of analyzing a correlation matrix which is complementary to CCA. Whereas CCA emphasizes global patterns, cluster analysis works "bottom up" by aggregating the most highly inter-correlated sets of variables first, and then working up to larger clusters which are less tightly inter-related. As a result, the first clusters identified the most likely sources of redundancy. As a side-benefit, the largest clusters allowed us to check the robustness of the CCA results (since cluster analysis and CCA are quite different mathematically).

Statistical significance

We used CCA as part of a spectrum of analytical tools. Therefore, it served to direct attention to patterns and to deviations from those patterns. We did not intend to place undue weight on the "statistical significance" of the CCA results. Nevertheless, we were interested in estimating the stability of the computed canonical correlations, and this required the calculation of standard errors.

The sampling theory for CCA is complex and assumes multivariate normality, an assumption far from the reality of our data: most of our dependent variables were discrete. Therefore, we turned to a well-known resampling method—the jackknife method—for estimating standard errors and confidence intervals (Efron & Tibshirani, 1993). We found the jackknife to be conceptually straightforward, though computationally demanding (see below).

The related problem of estimating the significance level of our canonical correlations called for a related solution—the use of randomization tests (Edgington, 1987). Resampling methods, such as the jackknife and the randomization test, are becoming more familiar and accepted; their detailed description is beyond the scope of this paper (see Peterson, 1991; Simon & Bruce, 1991).

We will, however, mention some of the computational problems deriving from our use of the jackknife and randomization methods. For complex analyses like CCA these resampling methods require fast computers and special techniques, as they necessitate the iterative solution of hundreds of matrix factorizations. Our programs were pieced together from library routines (Koeckler, 1994) and integrated using the Icon Programming Language (Griswold & Griswold, 1996), a high-level interpreted language. We also used a great deal of time-consuming check code. A jackknife analysis of 897 cases and 50 variables ran on a Pentium-based laptop in a little over three hours.

Interpretation and Visualization

Our final problem concerned interpretation of the results. We tried to find graphical methods which would help us understand and explain the multidimensional patterns found by CCA. These patterns are important because they help the analyst define, in a data-driven way, the most important environmental conditions and their corresponding effects on human action. One of the most helpful suggestions we found was due to Cliff (1987), who suggests interpreting structure correlations rather than weights. Structure correlations are the correlations of the X canonical variate with each of the original independent variables, and of the Y canonical variate with each of the original dependent variables. In this way, the somewhat mysterious canonical variates can be interpreted in terms of their correlations with the original variables. We then used two graphical methods to depict the pattern of structure correlations and to highlight deviations from the pattern and outliers (see Figure 1).

DISCUSSION AND CONCLUSIONS

CCA is the method of choice when dealing with multivariate dependent variables in a context otherwise suitable for multiple regression. CCA is best used as part of a suite of analytic methods. The full suite should include cluster analysis, state-transition (Markov) and dynamic modeling, graphical methods, and other statistical methods (Degani, 1996; Degani, Shafto, & Kirlik, 1995; Degani & Kirlik, 1995). Resampling methods may be used to compute confidence intervals and significance levels of canonical correlations. Structural correlations are helpful in interpreting the results of CCA, and simple graphical techniques can be used to understand and explain the results. CCA is capable of describing in an objective, data-driven way some of the complex patterns in data from field studies, simulations, and controlled experiments on human-machine interaction. It directs the analyst's attention to the main patterns in the data, as well as to the important deviations from these patterns.

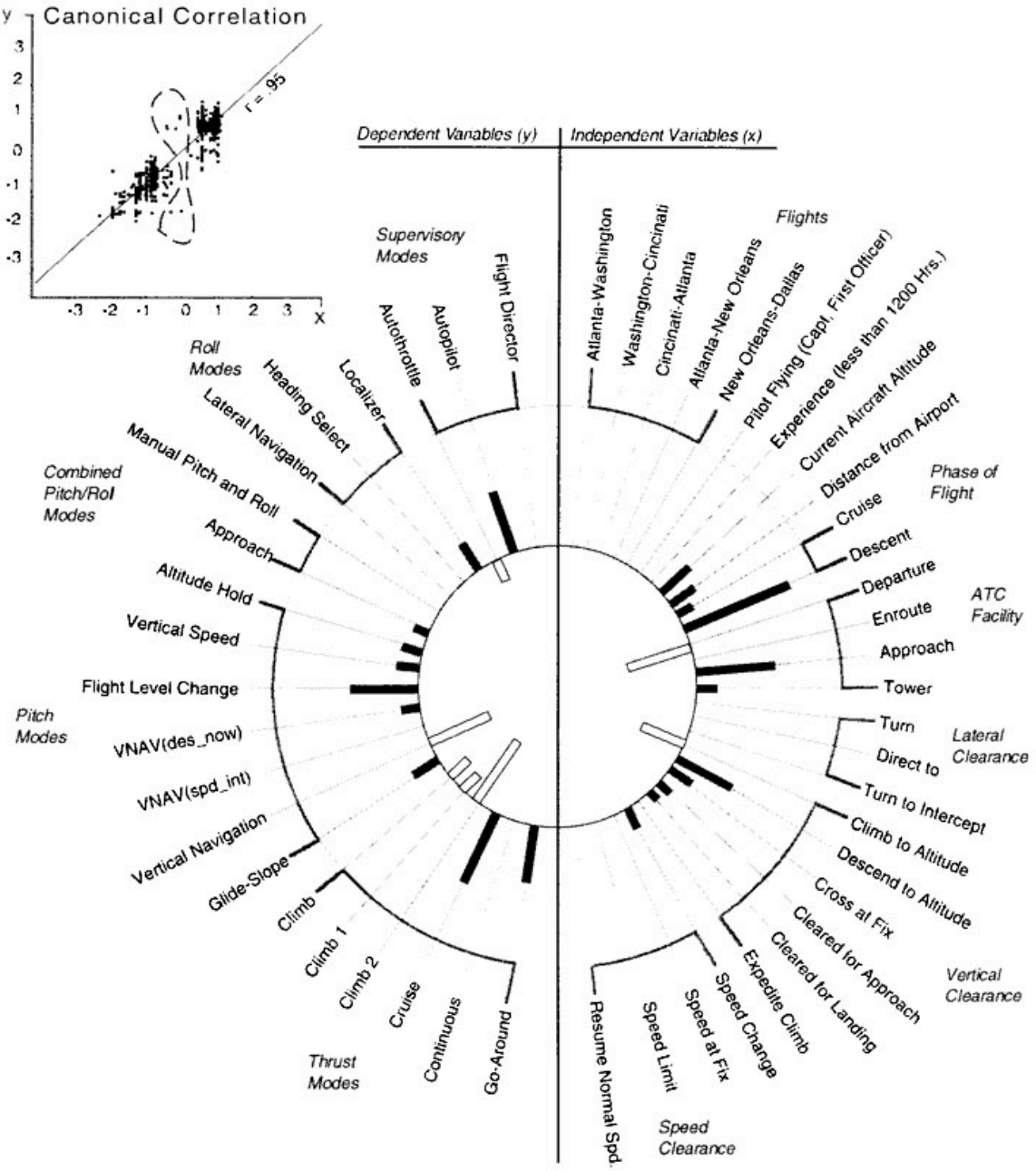

Figure 1. “Helio” chart

The relative size of structure correlations is indicated by the relative length of the bars extending toward the circumference (positive correlations) or toward the center (negative correlations). This very strong pattern indicates flight conditions (X variables) of descent with a “descend to altitude” clearance from Approach Control. This pattern is correlated with a response pattern (Y variables) of autopilot, “Flight Level Change” (FLCH) mode, cruise, and “Go-Around” mode armed; and “vertical Navigation (VNAV) mode and Thrust mode “Climb” disengaged. The plot in the upper left corner of the graph is the corresponding bi-variate canonical correlation. The relationship between the two composite variables (X and Y), plotted here in standard units, suggests a strong overall pattern (r=0.95). The few outliers indicated by the dotted lines need to be examined individually.

REFERENCES

Cliff, N. (1987). *Analyzing multivariate data*. San Diego: Harcourt Brace Jovanovich.

Degani, A. (1996). *Modeling human-machine systems: On modes, error, and patterns of interaction*. Unpublished doctoral dissertation. Atlanta, GA: Georgia Institute of Technology.

Degani, A., and Kirlik, A. (1995). Modes in human-automation interaction: Initial observations about a modeling approach. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics Conference* (pp. 3443-3450). Vancouver, British Columbia, Canada.

Degani, A., Shafto, M., and Kirlik, A. (1995). Mode usage in automated cockpits: Some initial observations. In T. B. Sheridan (Ed.), *Proceedings of the International Federation of Automatic Control; Man-Machine Systems (IFAC-MMS) Conference* (pp. 401-407). Boston, MA: IFAC.

Edgington, E.S. (1987). *Randomization tests* (second edition). New York: Marcel Dekker.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Griswold, R.E., & Griswold, M.T. (1996). *The Icon programming language* (third edition). San Jose, CA: Peer-to-Peer Publications.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology, 26*, 139-142.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. New York: Wiley.

Koeckler, N. (1994). *Numerical methods and scientific computing: Using software libraries for problem solving*. Oxford, UK: Clarendon Press.

McKeon, J.J. (1965). Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. *Psychometric Monographs, No. 13*.

Simon, J.L., & Bruce, P. (1991). *Probability and statistics the resampling way*. College Park, MD: University of Maryland.